DOCUMENT RESUME

ED 206 726 TM 810 638

AUTHOR Anderson, Scarvia B.
TITLE Testing and Coaching.
PUB DATE 15 Feb 31
NOTE 18p.: Paper presented at the Annual Meeting of the American Association of School Administrators (Atlanta, GA, February 13-16, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Aptitude Tests; Criterion Referenced Tests; Essay Tests; *Knowledge Level; Norm Referenced Tests; Objective Tests; Standardized Tests; Teacher Made Tests; Test Anxiety; *Test Coaching; *Testing; *Testing Problems; *Test Wiseness

ABSTRACT

This paper summarizes some of the issues involved in educational testing. First, it describes four of the fundamental characteristics of tests: whether they are measures of aptitude or achievement, whether they require students to produce or just recognize the best answer, whether they are interpreted in relationship to the scores of other students or in terms of some judgmental standard, and whether they are constructed by teachers in the schools or by some outside agency. Next, it reviews factors contributing to students' test performance, starting with their competence in the areas that the test measures. Other factors include the ability to follow directions and pace oneself, neatness and carefulness, and physical and emotional conditions. Finally, the current controversy about coaching for tests is examined, and the complexity of the issue is stressed. (Author/BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Presenter: Scarvia B. Anders
Title: Testing and Coaching
Date: February 15, 1981
Time: 3:30 p.m.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

S. Anderson and the Nat. School Boards Assn.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED206726

TM 810 638

Testing and Coaching

Scarvia B. Anderson
Educational Testing Service

We have a long history of explaining away any dissatisfaction with the educational "system" by blaming successively vulnerable parties: first we blamed the students, then we blamed their parents (indicting Dr. Spock along the way), then we blamed the teachers, and of course administrators are always blamed for everything. But the late seventies brought more bad news. SAT scores continued to decline, volunteers in the armed forces were labeled incompetent in the basic skills, colleges were required to offer more remedial courses, the crime rate went up. All of this in spite of unprecedented amounts of educational legislation and innovation. It was clearly time to nominate a new villain for the "what's-wrong-with-our-schools?" drama, and no group was more anxious to serve as the casting agent than those who had most recently been featured in the role--the teachers, or more specifically, one of the two major teachers' unions. Without so much as an audition, the NEA quickly picked standardized testing. The choice was praised by the NAACP, Ralph Nader, several state legislators, and two social scientists at the Harvard Medical School.

Their arguments were patterned more on the logic of Lewis Carroll, author of Alice in Wonderland, than of Charles Dodgson, mathematician. If minorities scored lower on the tests, if test scores were going down, if students' scores on tests could be improved, then the trouble lay with the tests and not with the schools. "If the tests were only cleared away," they said, "it would be grand."

(Reprinted with permission from the American School Board Journal, July 1981. Copyright © by the National School Boards Association. All rights reserved.)

Tests and Measurement

What are these tests that have assumed such prominence in the educational and political scene? Before I go on to talk about why students do well or poorly on tests and how we can help them do better, let's make sure we have some common understandings about the tests and the educational measurement process. How many of you have taken a formal course in tests and measurement? Well, even if you have, I am sure that it did not reduce most testing to the simple four-way treatment that I'm about to give it. Almost all educational tests can be characterized by whether they're intended as measures of aptitude or achievement, by what kind of response the student makes, by how the results are interpreted, and by who makes them up.

Aptitude-Achievement

There is a difference between aptitude and achievement tests, but it is not always striking. Performance on both of them depends on things you've learned, not just what you came into this world with. A wolf child would not do very well on these tests. For example, question 1 here is from a well-known measure of scholastic aptitude, while question 2 is from a basic skills tests that is usually called an "achievement test." [Visual 1] In general, so-called aptitude tests measure skills that have developed over a long period of time or applications of skills in new situations. It would be unusual to look at an aptitude-test question and be able to say precisely, "Oh, I learned that on Tuesday" or "Mrs. Parker taught me that." Aptitude tests are valued primarily for their predictive value. We say that scores on this test provide some indication of how well people will do in college, police academy, or advanced math. Since the skills measured on aptitude tests usually develop over a long period of time and from a variety of influences, it is usually not fair to blame the schools entirely if students'

aptitude test scores are low. This was what was done in many quarters when the SAT score decline became big news. I admit that you would expect the schools to shoulder some blame if, after 11 or 12 years of "exposure," students couldn't handle words and number concepts as well as students of a few years before. But the SAT did not measure a lot of learnings the schools said they were trying to foster, the population of students sitting for the SAT changed somewhat, and many other societal influences were operating.

Achievement tests are generally more easily understood and appreciated by the public than aptitude tests, because the questions on achievement tests can be more readily related to the school curriculum. Scores on tests of science, history, literature, and mathematics can be expected to be influenced by the quality and recency of relevant instruction as well as by the student's inclination to learn. It is patently fair to blame the school and the teacher--as well as the students--if students do badly on an appropriate test over the course they have just completed or in a skill in which they've been trained daily for several years.

To oversimplify, the information from aptitude tests is "looking ahead" information. Does the student have the base necessary to help him succeed in the future? The information from achievement tests is "looking back" information. Did the student learn thus and so? Of course, if students didn't learn what the schools attempted to teach, and schools find this out in a reasonable time, they can institute some procedure to try to remedy the situation. (One hears fewer charges of test bias levied against achievement tests than against aptitude tests.)

Most of the minimum competency tests that are around are somewhere between aptitude and achievement measures, as they have been thought of conventionally. Competency tests measure skills which, for some students anyway, have been acquired over a rather long period of time, and these

tests sometimes include applications of skills to new materials; on the other hand, they measure skills that the schools have always had a clear directive to cultivate and are held almost fully responsible for.

Recognition-Production

A second major way in which educational tests vary is in the kinds of responses students are required to make. Responses are basically of two kinds: The student has to recognize the appropriate answer to a question or problem, or the student has to produce an answer. The recognition questions on most professionally produced tests are multiple-choice items. Teachers frequently use true-false or matching items as well. In spite of popular belief to the contrary, multiple-choice items can provide effective measurement of reasoning processes as well as straight memory. Look at question 3. [Visual 2] The question does not simply ask: What areas of the world have the highest population densities? Rather, it presents a novel situation in which the student must infer that, of the choices offered, only population density (A) provides a plausible explanation of the shadings on the map.

The most frequently used "production" tests are short-answer or "fill-in-the-blank" tests and essay examinations. An example of a short-answer test item: The sum of scores divided by the number of scores is called the _____. The student would write in mean, arithmetic mean, or average. You are all very familiar with essay questions from your own school days. For example: What is meant by the statement that France, before 1789, was centralized but not unified? Or: Why did the Dreyfus affair become an issue of national significance? Or, more typically: Describe your summer vacation. Of course, the better constructed the questions and the clearer the directions to students, the easier essay tests are to

score. But, even then, scoring is not simple. To avoid biased and idiosyncratic scoring, scorers need to have a clear outline in advance of just what components will be scored and how much weight will be given to each. For example, will essays be scored for style and mechanics as well as substance? A successful scoring technique used by professional test makers, but rarely by teachers, is called "holistic scoring." Raters agree on the characteristics that a top paper will have and choose samples that they think epitomize that standard. Then they do the same for papers to receive each of the lower ranks. The top papers might be scored 5, the poorest papers 1. Then scores are assigned to all of the papers by comparing them to the samples that have been designated.

In the face of the scoring problems that production tests present, as well as the fact that fewer such questions can usually be presented in a given period of time--and thus less comprehensive coverage of an area is possible--you may wonder why testers use such questions at all. The choice of production rather than recognition questions depends directly on the purpose of the test. If you want to find out how well students write, then there's no real substitute for asking them to write. To ask them questions <u>about</u> writing is not the same thing. You can think of many other examples: translating a passage from Cicero, speaking extemporaneously on a popular topic for five minutes, analyzing a blood sample, editing a technical manuscript, making a soufflé, disassembling or assembling a rifle.

This distinction between recognition and production tests is an important one. In many cases, people talk as if they are testing one way when they are really testing another. A student who may recognize appropriate laboratory procedures, say, from word and picture descriptions might very well add water to sulphuric acid, instead of vice versa, when left on her own in the lab.

Interpretation

A third major dimension of tests is how the scores are interpreted. There are two main ways: by reference to the performance of other people and by reference to judgmental standards. Tests interpreted in the first way are referred to as "norms-referenced," and tests interpreted in the second way are called "criterion-referenced." In the first instance, we would say, for example, that Betty's score is in the top quarter of scores of students in her graduating class, or above the mean for a national group of 12th graders, or at the 60th percentile for seniors in the state (meaning that 60 percent of the seniors in the norms group scored lower than she did). This example also indicates that, with respect to the skills measured, Betty had more competition in the state than in her own class, for she stood higher in her class than she did in the state as a whole.

Criterion-referenced tests are interpreted according to judgments that have been made about what constitutes appropriate performance--highly satisfactory, adequate, or minimal--depending on the test and the purpose for which it was given. For example, a group of teachers might agree that 60 percent was a passing score on Test X or that 10 out of 12 items on a math subtest indicated substantial mastery of the concept covered. An important thing to keep in mind is that there are no "automatic" criteria around for school tests. Sixty percent might be a passing score for one test but not considered passing for another easier test. Ten out of 12 items might be considered "mastery" for one concept but not for another.

It is unfortunate that criterion- and norms-referenced measurement are viewed in sharp contradistinction by some measurement cults--and sometimes with the connotation of "good" versus "bad" practice. In fact, the two approaches frequently overlap and are certainly useful supplements to each other. Moreover, normative considerations usually underlie choices of

measurement content and standards for criterion-referenced tests, although these considerations may be implicit rather than explicit. For example, we may set 50 words per minute with two errors as the minimum standard for graduation from a 3-month typing course. We don't set 10 w.p.m. or 150 w.p.m. because these speeds do not reflect typical typing requirements or average human capability--both normative notions.

Whether test users rely more heavily on norms- or on criterion-referenced interpretations of test performance, they still must rely to a great extent on judgment in setting standards or cut-off scores. For example, in a particular location characterized by a teacher surplus the accrediting agency might qualify only those applicants who score above the 25th percentile. One of the mysterious phenomena you must run into is that many teachers and parents take the 50th percentile or mean as their implicit standard of adequate performance. They will bewail the fact that Maria or Fred is below average without stopping to think that half the children have to be. In fact, it is a mistake to think of "average" as representing a single score. It is better to think of "average" as a range of scores encompassing the middle third of the group.

There are many different procedures that educators use to set standards or cut-off scores. These vary in formality and elaborateness. For the moment, the message I want to leave with you is that the burden is on educators to make the standards they adopt explicit and to provide a clear rationale for them. This is of paramount importance if cutting scores are to be used for such crucial decisions as graduation, promotion, or admission to a program of study.

Test Developers

The fourth dimension of tests to consider is who makes them up--teachers or some outside agency. By far the largest number of tests used are those given by individual instructors. The critics of educational

testing usually overlook this fact. For every SAT, there are thousands of mid-term and final exams in mathematics and English. For every standardized elementary reading test, thousands of dittoed reading exercises used in classrooms. It is classroom tests and other evaluations by instructors that are, and will continue to be, major factors in promotion, graduation, recommendations for further education and employment, and students' self-appraisals of their academic "fitness."

The two major kinds of outside groups that develop and supply tests are state departments of education and national testing agencies such as ETS; ACT (American College Testing Program); Psychological Corporation; Harcourt, Brace, Jovanovich; Houghton-Mifflin; and CTB/McGraw-Hill. Sometimes the states and national agencies work together in contractual relationships, especially when state departments of education have limited technical or technological resources.

In the case of achievement measures, ideally those developed and used to evaluate students and guide their learning at the local level and those selected or imposed from outside sources should be complementary. A mandated state basic skills test should not come as a shock to a student. If the state test is an appropriate measure of school learning, then it should look not unlike the exercises the student has already met--frequently--in his or her own school. A lack of congruency is a legitimate source of concern to educational critics. Either the school has not prepared students properly or the external test is an unfair reflection of school objectives.

In the case of aptitude tests, we would not expect the external tests to bear as close a relationship to the achievement tests normally used in schools. You will recall that aptitude tests usually measure knowledge and skills accumulated over a relatively long period of time and from many sources. For example, a student might encounter a vocabulary word he has never seen before. What we would hope, however, is that

from an accumulated knowledge about words and how they are used the student might be able to deduce the meaning from the structure of the word and from the context.

This is the end of my mini-course in tests and measurements. I hope that you are suitably impressed by the number and variety of tests that can be described by the four two-way classifications. There are criterion-referenced achievement tests that are built by teachers and require students to produce the answers--the category that probably characterizes the largest number of educational tests given and taken in the world. There are criterion-referenced multiple-choice tests that are constructed by state departments of education to measure achievement in the basic skills. There are norms-referenced, multiple-choice, aptitude tests developed by agencies such as ETS, usually for some group or association such as the College Board. These last tests are the ones that have been of greatest concern to school people in recent years, probably because they are the tests over which school people have the least direct control.

Factors in Test Performance

Let us consider for a moment the factors that contribute to a student's performance on any kind of test, teacher-made or standardized. If the test is a good test--valid and reliable are the terms test specialists use--then the student's score is determined first and foremost by whether he or she knows the answer to the questions on the test or can do the tasks that the test requires. We must remember this at all times: There is no magic that will enable students to succeed on a good test if they aren't up on the subject matter or skills that the test measures. (Short of cheating, that is.)

I'll tell you a story about the director of research in a large public school system. He was dismayed because the students' scores on a national standardized reading test had dropped from the year before, and he wouldn't release them--to the superintendent, the press, or anybody. First, he got the computer center to rescore the tests--he was sure there was an error. There wasn't. Then he had them check on all the statistical computations. Still OK. Finally, after several weeks of this and clamors on all sides to see the scores, the head of the computing center met the research director in the parking lot. "Hey, buddy," he said, "I think I've figured out how to get those test scores up." "How? How?" insisted the research director, with relief showing on his face for the first time in weeks. "Teach those kids how to read," was the answer.

I am frequently asked if tests are biased--without any specification of the kind of test or the kind of students or what the questioner means by "biased." Sometimes at the end of a long day or my patience, I snap back, "Sure, they're biased--against those who can't read or spell or do whatever the test is about!"

Given that test performance is determined largely by what students know or can do, what are some other possible, if lesser, influences on test scores?

1. One is certainly whether they understand and follow the test directions. If the directions say to select the <u>least</u> probable explanation for an event and they select the most probable explanation, then they're obviously going to miss the question. If one of the choices for a set of math items is "The answer cannot be determined from the information given" and they ignore that choice, then their scores will suffer. If they're told to answer any four out of five essay questions, and they answer all five, then they may have wasted time that they could have devoted to more complete

answers to four questions. And woe be to the test taker who skips over to part 2 when he or she is supposed to be working on part 1 of a test!

2. A second skill that may affect performance when there are several questions on a test and limited time to answer them is pacing. Students should pace their work so that they have sufficient time to get through the questions they know and don't get hung up on the questions they're unsure of. They should save them to ponder over at the end.

3. Third, neatness and carefulness can count. An essay can be brilliant, but the brilliance may not shine through a sloppy or illegible paper. Attention to clerical detail--matching question numbers to spaces on the answer sheet, erasing initial answers when students change their minds--can be of major importance on a machine-scored test.

4 and 5. Finally, physical conditions and emotional states can contribute to performance on a test. In both cases, there is probably a rather broad range of tolerance before performance will be adversely affected. A little cold probably won't have a noticeable effect on test performance, but a bad case of the flu with accompanying fever could seriously reduce a student's ability to attend to the test tasks. Similarly, rather poor lighting or a testing room that is a little too warm might make the test taker uncomfortable but would be unlikely to jeopardize his or her test performance if the student is capable of doing the tasks on the test. One of my professors described studies he did in World War II that showed that aptitude test scores were little affected even when recruits took the test right after an all-night march.

Of course, serious emotional disturbances can affect test performance. The most directly relevant of these is described by the term "test anxiety." We know that some people get more anxious about a test than others and that anxiety varies with how well prepared students think they are and the cruciality of the testing situation. For example, we would expect students

to be more anxious on a test that was to be used in determining their eligibility for a significant and much needed award than on a test that was only one of several in a given school course. (However, we shouldn't forget that there are a few among us who actually like taking tests and are inclined to "give their all" in a testing situation regardless of the purposes of the test.) The measurement and study of test anxiety are complex, and have challenged a number of prominent psychologists through the years. In spite of all the work, we are not in a position to predict just how many people in a given population will have such severe test anxiety that it will be paralyzing, but the number is probably quite small. On the other hand, we know that a little anxiety is probably facilitating rather than debilitating. (Think about your own performance in situations where you know you are going to be judged by important audiences.) Even though I believe that test anxiety is not a serious determiner of performance for very many people, I would not dismiss the value of the concept as an ego-saving device for thousands of students and their parents. If the term "test anxiety" did not exist, we would probably have to invent it or something like it to preserve the self-concepts of those who might otherwise have to turn to such alternative explanations as their own ignorance or laziness about studying for poor test performance.

Now, I have talked about six influences on test scores: abilities and knowledge required by the test, understanding and following directions, pacing, neatness and carefulness, physical state and conditions, and emotional state. What can we as educators and testers do to minimize negative influences in these areas?

1. We can certainly design testing conditions so that directions are as clear as possible, test administrators are friendly and helpful, the physical environment is comfortable, and students who are ill are offered a make-up exam.

2. We can also teach students test-taking skills--pacing, handling machine-scored answer sheets, understanding different types of questions, and so on. Differential familiarity with objective tests is not the problem it once was. In the first testing study I ever did--back in 1951 at George Peabody College for Teachers--we found that many of the freshmen had had no previous experience with objective tests and machine-scored answer sheets, and we instituted practice tests before the real tests in order to try to even out any differences associated with the mechanics of test taking. Today almost all children are exposed to a wide variety of standardized tests as they go through school.

3. If we can identify the most extremely anxious students, we may be able to recommend personal therapy for them.

What can we do, though, about the biggest factor that affects students' scores--their knowledge and skills? The best answer to that was contained in the anecdote I told you a few minutes ago--teach them to read or write or whatever the test demands. Unfortunately, if the knowledge or skills are significant, this usually takes a rather long time. So everyone wants to know if there are any short-cuts.

Coaching

One of the most confusing arguments about standardized tests has to do with whether they are "coachable." The NEA, Ralph Nader, Slack and Porter say that tests are no good if they are. The first thing that this argument denies is the link between education and the testing process. Personally, I think that there is little place in education for tests where performance cannot be influenced by instruction. But to discuss coaching intelligently we need first to consider the nature of the coaching, the length of the coaching, and the kind of test that is being coached for. I am sorry to burden you with all of these conditions, but, if you leave this room with only one thought,

I want it to be that coaching is a complex issue that is not satisfied by "off-the-cuff" answers.

I have already mentioned training in test-taking skills, which is a significant part of most coaching efforts. Now let me concentrate on coaching in the skills or knowledge measured by a test.

"Coaching" can usually be distinguished from regular instruction by its duration, purposes, and techniques:

- Instruction is generally long-term; coaching generally involves a few hours, days, or weeks.
- Most people will agree that the purpose of instruction is to develop skills, knowledge, and understandings to an appropriate level of mastery--for their own sake; presumably students would also do better on tests measuring the skills, knowledge, and understandings involved. The basic purpose of coaching is to improve performance on a particular examination; if students also improve in the relevant skills, knowledge, and understandings, well and good.
- A good instructional program generally includes alternative explanations of phenomena and solutions to problems, practice with variations, and a certain amount of enriching detail --all designed to help the student understand as well as know. On the other hand, a coaching effort concentrates only on the kinds of questions that are likely to be on the test, quick solutions to typical problems, and associations of test language with certain types of responses. Drill is a significant part of most coaching efforts.

I was "coached" in the evenings for three weeks for my PhD examination in German. Since the examination would require only translation and students

were allowed to use a dictionary, I did not spend much time learning long lists of vocabulary words. I did memorize the articles and common prepositions, conjunctions, adjectives, and adverbs so I wouldn't have to spend time looking them up. The tutor concentrated on teaching me how to look quickly at a German sentence and pick out the subject and predicate and what words went with each. And, since we knew the exam would be in my field, we practiced on abstracts of articles from German psychological journals, not with travel books or short stories. If my object had been learning how to read scientific German fluently so I could keep up with the literature, we would obviously have taken a very different approach and much longer time. And, if I had really wanted to learn German, even more would have been required.

I passed the exam handily, but know practically no German today. Unfortunately, retention from coaching efforts is frequently poor. Coaching seems to work best when it is of the "refresher" variety. If a student has been out of school for a few years or hasn't studied math recently, then coaching can clear away some of the rustiness from skills which the student had.

The amount of student contact time in a coaching program seems to make a difference, up to a point. Studies with the SAT show that there is a positive relationship between amount of contact time and score increases associated with coaching. However, the law of diminishing returns soon begins to operate. For example, an increase of 10 points on the 200- to 800-point verbal scale might be associated with 12 hours of coaching, but 20 points would require 57 hours, and 30 points 260 hours. Extrapolating beyond existing data, it appears that for a 40-point gain something close to 1200 hours of contact time would be required--or full-time schooling!

Most of the public discussion of coaching has pertained to aptitude tests. If we recall the continuum from aptitude tests (or tests of general abilities) to specific achievement tests, we can see that the broader and more general the test content the more difficult it is to "cover" it in a short time. For example, it would be impossible in a couple of weeks to anticipate or teach all the words that might be in a general vocabulary test for college entrants. But, if the test were on American literature, one could certainly hit some of the high spots in two weeks. And, if the test were to be based on an even more narrowly circumscribed content (for example, symbols for the elements), one might learn all of the items that might be included in a relatively short period of time.

To summarize my remarks and position on coaching:

1. Those who would condemn tests because they are coachable ignore a fundamental relationship between educational testing and teaching.

2. The success of coaching depends on many factors including the student's motivation, the nature and duration of the coaching effort, and the kind of test involved. In general, aptitude tests are harder to coach for than achievement tests.

3. Whether or not coaching produces actual score increases in particular instances, many students feel that it is helpful and cite reductions in test anxiety as a primary reason. Therefore, schools that have sufficient numbers of students taking tests considered of major importance would be well advised to provide those students with appropriate opportunities to prepare for the tests--in both the substance and the mechanics of taking them. This would help avoid charges of inequity between students who can afford expensive commercial coaching courses and students who can't.

Summary

I have ranged very widely today over some of the issues involved in educational testing. First I lectured you on four of the fundamental

characteristics of tests: whether they are measures of aptitude or achievement, whether they require students to produce or just recognize the best answer, whether they are interpreted in relationship to the scores of other students or in terms of some judgmental standard, and whether they are constructed by teachers in the schools or by some outside agency. Don't forget that the largest number of tests given in the world today are made up by teachers.

Then we reviewed some of the factors that contribute to students' test performance, starting with their competence in the areas that the test measures. Beyond that, we recognized that ability to follow directions and pace oneself appropriately, neatness and carefulness, and physical and emotional conditions can influence test scores. We noted, however, that human beings have a rather broad tolerance with respect to both physical and emotional conditions, and usually only major variations from normal will affect test scores.

Finally, we examined the current controversy about coaching for tests, and I hope I dispelled any notions that coaching is a simple issue with a single solution.

I began my remarks with a statement about the shifting blame that we have seen over the last fifty years for "what's wrong with our schools." The tests are the latest culprit, but you and I are realists enough to know that that charge is not going to stick. As Albert Shanker put it, "You can't blame the thermometer for the patient's fever." With the growing public and legislative demands for educational accountability and productivity, tests will probably occupy a more rather than less important place in the educational scene. Thus, it behooves you as administrators to learn as much about measurement as you can, so that you can monitor testing practices in your schools, resist inappropriate impositions of tests on your systems, counter the misuse of tests in unwarranted attacks on education, and, most important, harness these important tools for your own decision-making needs.